~~PDF~~ Content Metadata

Ross Mounce

University of Bath, PhD Candidate
Open Knowledge Foundation,
Panton Fellow & Community Coordinator, Open Science

(I have many 'hats' these are just some)

Open Knowledge Foundation

FYI, this is also on: slideshare

@rmounce
#btpdf2

UNIVERSITY OF BATH

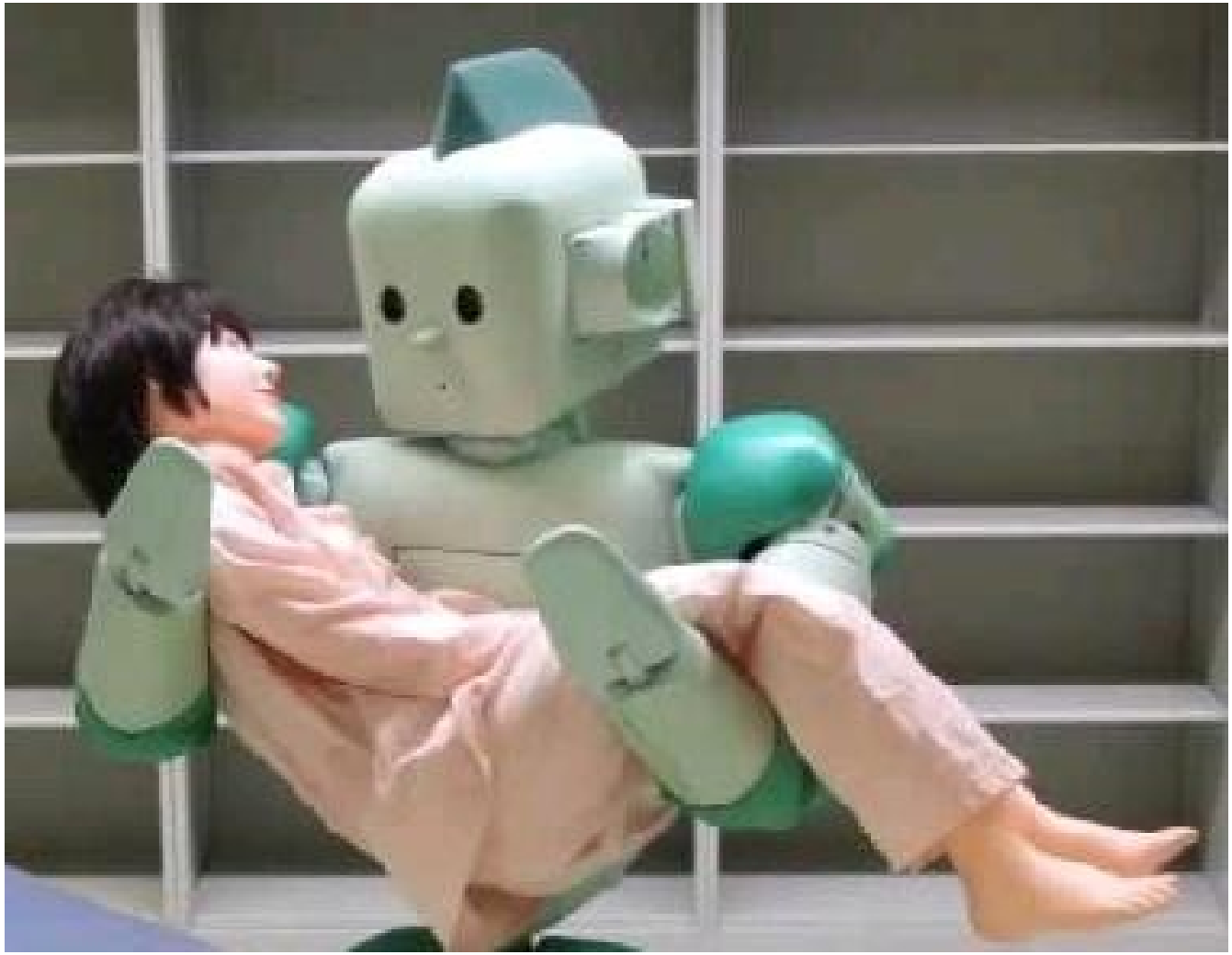
Why is metadata important?

- Millions of scholarly articles published per year

What is in 'sdarticle.pdf' & what is in 'sdarticle (1).pdf'

I don't want to have to read PDFs myself to know roughly what's inside!!!

Machines can help us



ExifTool reads embedded metadata

- Quick & easy to get the metadata of all files

ISO 16684-1:2012 standard
XMP -Extensible Metadata Platform

www.sno.phy.queensu.ca/~phil/exiftool/

...in practice PDF metadata is poor

Some publishers do not have **any** embedded XMP in their 'Version of Record' PDFs (e.g. Taylor & Francis from what I've seen so far)

...and I've yet to see a PDF 'in the wild' with embedded **license information**. Very important! [If I downloaded these files years ago are they CC BY, 'All Rights Reserved' or CC BY-NC-ND ?]

Data on

figshare
credit for **all** your research

://rossmounce.co.uk/2013/01/06/pdf-metadata-using-exiftool/
://dx.doi.org/10.6084/m9.figshare.106195

://rossmounce.co.uk/2012/12/31/pdf-metadata-why-so-poor/
http://dx.doi.org/10.6084/m9.figshare.105633

What producer wouldn't properly label their content?

Label your products!

If I can't redistribute it, put that in the metadata please!

Can you tell me anything *useful* about this PDF from its metadata?

[XMP] XMP Toolkit : Adobe XMP Core 4.2.1-c043 52.389687, 2009/06/02-13:20:35
[XMP] Producer : Acrobat Distiller 6.0 (Windows)
[XMP] Create Date : 2008:07:16 14:00:37+05:30
[XMP] Creator Tool : PScript5.dll Version 5.2.2
[XMP] Modify Date : 2010:10:11 08:21:42-07:00
[XMP] Metadata Date : 2010:10:11 08:21:42-07:00
[XMP] Format : application/pdf
[XMP] Creator : Administrator
[XMP] Title : ST&HV306704.qxd
[XMP] Document ID : uuid:6a0d428a-3141-43c3-bdbe-8446710f5c8e
[XMP] Instance ID : uuid:f7caddbd-1dd1-11b2-0a00-15f6e8a599ff

PDF from http://dx.doi.org/10.1177/0162243907306704 (published 2008)